Statistical Methods & Applications manuscript No.
(will be inserted by the editor)

On Wald Tests for Differential Item Functioning Detection

Received: date / Accepted: date

Abstract Wald-type tests are a common procedure for DIF detection among the IRT-based methods. However, the empirical type I error rate of these tests departs from the significance level. In this paper, two reasons that explain this discrepancy will be discussed and a new procedure will be proposed. The first reason is related to the equating coefficients used to convert the item parameters to a common scale, as they are treated as known constants whereas they are estimated. The second reason is related to the parameterization used to estimate the item parameters, which is different from the usual IRT parameterization. Since the item parameters in the usual IRT parameterization are obtained in a second step, the corresponding covariance matrix is approximated using the delta method. The proposal of this article is to account for the estimation of the equating coefficients treating them as random variables and to use the untransformed (i.e. not reparameterized) item parameters in the computation of the test statistic. A simulation study is presented to compare the performance of this new proposal with the currently used procedure. Results show that the new proposal gives type I error rates closer to the significance level.

Keywords Differential item functioning · False positive rate · Item response theory · Lord test · Type I error rate · Wald test

Mathematics Subject Classification (2000) MSC 62 Statistics

1 Introduction

Item Response Theory (IRT) provides a framework for the statistical analysis of the responses given to the items of a test or questionnaire (Bartholomew et al, 2011; van der Linden, 2016). In IRT models, the probability of observing a certain response to an item is modelled as a function of a latent variable

and some parameters related to the item. These models, originally developed for the assessment of learning outcomes, now find application in many other contexts, including medicine and psychology (Reise and Revicki, 2014). Differential Item Functioning (DIF) is a violation of the invariance assumption of IRT models, and occurs when the probability of a positive response for examinees at the same ability level varies in different groups. Various methods have been proposed in the literature for the detection of DIF (see for example Magis et al, 2010). Among them, the Lord's chi-square test (Lord, 1980) is a common procedure that presents the advantage of requiring the estimation of the item parameters just ones for each group, as the selection of the anchor items (which are the items free of DIF) is performed in a second step. The test was originally developed for detecting DIF between two groups, and then extended to the case of multiple groups by Kim et al (1995). However, simulation studies reported in the literature showed that the empirical type I error rates for this test departs from the significance level (Kim et al, 1994). In particular, they are largely greater than the significance level for the Three-Parameter Logistic (3PL) model, while they are smaller for the Two-Parameter Logistic (2PL) model. In this paper, the reasons of this discrepancy will be discussed, and a new proposal will be presented. The new proposal applies to multiple groups as well as to two groups.

This paper is structured as follows. Section 2 reviews IRT modeling, the traditional procedure currently used for the Wald test and its extension to multiple groups. Section 3 illustrates the new proposal, which is compared to the traditional procedure by means of simulation studies in Section 4. Finally, Section 5 contains some concluding remarks.

2 Preliminaries

2.1 IRT modeling

Let Y_{ij} be the dichotomous response given by subject i to item j, where 1 denotes a correct response and 0 denotes an incorrect one. In a 3PL model, the probability of observing a response equal to 1 to item j, given the latent variable θ_i, is given by

$$p_{ij} = p(Y_{ij} = 1|\theta_i; a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp\{Da_j(\theta_i - b_j)\}}{1 + \exp\{Da_j(\theta_i - b_j)\}}, \qquad (1)$$

where a_j, b_j and c_j are the discrimination, difficulty and guessing parameters, while D is a constant typically set to 1.7. The 2PL model is obtained when the guessing parameters c_j are set to 0, while the Rasch model requires also that the discrimination parameters are equal to 1. The item parameters are generally estimated by means of the marginal maximum likelihood method (Bock and Aitkin, 1981). This approach assumes that the latent variable follows a standard normal distribution, and the marginal distribution of the responses given by one subject is obtained by integrating the joint probability over θ.

While IRT models are generally specified as in Equation (1), the parameterization used for estimation is as follows (Bock and Aitkin, 1981; Patz and Junker, 1999; Rizopoulos, 2006)

$$p_{ij} = P(Y_{ij} = 1|\theta_i; \beta_{1j}, \beta_{2j}, \beta_{3j}) = c_j + (1 - c_j)\frac{\exp(\beta_{1j} + \beta_{2j}\theta_i)}{1 + \exp(\beta_{1j} + \beta_{2j}\theta_i)}, \quad (2)$$

with

$$c_j = \frac{\exp(\beta_{3j})}{1 + \exp(\beta_{3j})}. \quad (3)$$

The set of parameters for each item is then $\{\beta_{1j}, \beta_{2j}, \beta_{3j}\}$, while the parameters of the usual IRT parameterization given in (1) are obtained after the estimation, using these transformations:

$$a_j = \frac{\beta_{2j}}{D}, \quad (4)$$

$$b_j = -\frac{\beta_{1j}}{\beta_{2j}} \quad (5)$$

and Equation (3). The estimation of the parameters requires the maximization of the marginal likelihood function, which is given by

$$L(\boldsymbol{\beta}) = \prod_i \int \prod_j p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}} \phi(\theta_i) d\theta_i, \quad (6)$$

where $\boldsymbol{\beta}$ is the vector containing the parameters of all the items, and $\phi(\cdot)$ denotes the density of the standard normal variable. Constraining the mean and the standard deviation of the latent variable to be equal to 0 and 1 is necessary to ensure identifiability of the parameters of the model (van der Linden, 2016, §2.2.3). A special case is given by the Rasch model, which requires only one constrain (typically the mean of the latent variable equal to 0). This can be accomplished in standard software that uses the standard normal for the latent variable by constraining the discrimination parameters to be constant, i.e. $\beta_{2j} = \beta_2$ in Equation (2). As a consequence of the constrains needed to assure identifiability, when the item parameters are estimated separately for different groups of subjects, they are not directly comparable, and require a linear transformation in order to obtain values expressed on a common scale (van der Linden, 2016, §2.2.4).

2.2 Wald tests for DIF

One of the fundamental assumptions of IRT models is the invariance of the item parameters. DIF is a violation of this assumption, and occurs when the item parameters have different values in different groups of subjects (Magis et al, 2010). The Lord's chi-square test is a Wald-type test based on the comparison of the item parameter estimates obtained from different groups. Let

$\mathbf{v}_{jk} = (a_{jk}, b_{jk}, c_{jk})^\top$ be the vector of item parameters for group k. The test was originally formulated for the case of two groups under investigation. The null hypothesis is the invariance of item parameters across groups

$$H_0 : \begin{pmatrix} a_{j1} \\ b_{j1} \\ c_{j1} \end{pmatrix} = \begin{pmatrix} a_{j2} \\ b_{j2} \\ c_{j2} \end{pmatrix}.$$

Without loss of generality, throughout this paper it is assumed that the reference group is group 1.

Before comparing item parameter estimates deriving from different groups, it is then necessary to transform them in order to obtain values expressed on the same metric. The equating transformations that permit to transform the item parameters estimates from the scale of group k to the scale of the reference group are

$$\hat{a}^*_{jk} = \frac{\hat{a}_{jk}}{A_k}, \tag{7}$$

and

$$\hat{b}^*_{jk} = A_k\, \hat{b}_{jk} + B_k, \tag{8}$$

where A_k and B_k are two constants called equating coefficients (Kolen and Brennan, 2014). The guessing parameters c_j do not need to be transformed. The test statistic is

$$\chi^2_j = (\mathbf{v}_{j1} - \mathbf{v}^*_{j2})^\top (\boldsymbol{\Sigma}_{j1} + \boldsymbol{\Sigma}^*_{j2})^{-1} (\mathbf{v}_{j1} - \mathbf{v}^*_{j2}), \tag{9}$$

where the vector of estimates of the parameters of item j in group k is

$$\mathbf{v}_{jk} = (\hat{a}_{jk}, \hat{b}_{jk}, \hat{c}_{jk})^\top,$$

the vector of estimates transformed to the scale of the reference group is

$$\mathbf{v}^*_{jk} = (\hat{a}^*_{jk}, \hat{b}^*_{jk}, \hat{c}_{jk})^\top,$$

$\boldsymbol{\Sigma}_{jk}$ is the estimated covariance matrix of $\mathbf{v}_{jk}$ and $\boldsymbol{\Sigma}^*_{jk}$ is the estimated covariance matrix of $\mathbf{v}^*_{jk}$.

Kim et al (1995) extended the test to the case of multiple groups, considering as null hypothesis

$$H_0 : \begin{pmatrix} a_{j1} \\ b_{j1} \\ c_{j1} \end{pmatrix} = \cdots = \begin{pmatrix} a_{jk} \\ b_{jk} \\ c_{jk} \end{pmatrix} = \cdots = \begin{pmatrix} a_{jK} \\ b_{jK} \\ c_{jK} \end{pmatrix} \tag{10}$$

and as test statistic

$$Q_j = (\mathbf{C}\mathbf{v}_j)^\top (\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^\top)^{-1} (\mathbf{C}\mathbf{v}_j), \tag{11}$$

where

$$\mathbf{v}_j = (\mathbf{v}_{j1}^\top, \mathbf{v}_{j2}^{*\top}, \ldots, \mathbf{v}_{jK}^{*\top})^\top,$$
$$\boldsymbol{\Sigma}_j = \mathsf{COV}(\mathbf{v}_j) = \text{blockdiag}(\boldsymbol{\Sigma}_{j1}, \boldsymbol{\Sigma}^*_{j2}, \ldots, \boldsymbol{\Sigma}^*_{jK}),$$

blockdiag$(\cdot)$ denotes a block diagonal matrix and $\mathbf{C}$ is a contrast matrix. When $K = 2$, Equation (11) returns the test statistic (9). Under the null hypothesis, the asymptotic distribution of the test statistic is a Chi-square distribution with degrees of freedom equal to the number of rows of the matrix $\mathbf{C}$.

3 A new proposal

Simulation studies reported in the literature (Kim et al, 1994) showed that the empirical type I error rate for this test diverges from the significance level. In particular, it is largely greater for the 3PL model, while it is smaller for the 2PL model.

The proposal of this paper aims at narrowing the discrepancy between the empirical type I error rate and the nominal value. Two issues will be considered to this end. First, the equating coefficients in Equation (7) and (8) are treated as known constants in the computation of $\boldsymbol{\Sigma}^*_{jk}$ (see for example Kim et al, 1994, 1995), while they are actually estimated. The literature on test equating provides various methods for the estimation of the equating coefficients (Kolen and Brennan, 2014), and the asymptotic standard errors are derived in Ogasawara (2000) and Ogasawara (2001). The proposal of this paper is to account for the estimation of the equating coefficients in the computation of the covariance matrix of the item parameters.

A second issue regards the parameterization usually used for the estimation of the parameters, which is given in Equation (2). The item parameters in Equation (1) are obtained in a second step, and the covariance matrices $\boldsymbol{\Sigma}_{jk}$ are obtained by applying the delta method (Casella and Berger, 2002, §5.5.4).

Of course, the item parameter estimates need to be converted to a common metric. This task can be performed using the following equations:

$$\hat{\beta}^*_{2jk} = \frac{\hat{\beta}_{2jk}}{\hat{A}_k}, \tag{12}$$

and

$$\hat{\beta}^*_{1jk} = \hat{\beta}_{1jk} - \hat{\beta}_{2jk}\frac{\hat{B}_k}{\hat{A}_k}. \tag{13}$$

The derivation is given in Appendix A.

The proposal of this paper is to compute the test statistic using untransformed item parameter estimates. The null hypothesis is then

$$H_0 : \begin{pmatrix}\beta_{1j1}\\ \beta_{2j1}\\ \beta_{3j1}\end{pmatrix} = \cdots = \begin{pmatrix}\beta_{1jk}\\ \beta_{2jk}\\ \beta_{3jk}\end{pmatrix} = \cdots = \begin{pmatrix}\beta_{1jK}\\ \beta_{2jK}\\ \beta_{3jK}\end{pmatrix}, \tag{14}$$

which is equivalent to (10). The test statistic is given by

$$W_j = (\mathbf{C}\boldsymbol{\beta}_j)^\top(\mathbf{C}\boldsymbol{\Omega}_j\mathbf{C}^\top)^{-1}(\mathbf{C}\boldsymbol{\beta}_j), \tag{15}$$

where

$$\boldsymbol{\beta}_j = (\boldsymbol{\beta}_{j1}^\top, {\boldsymbol{\beta}^*_{j2}}^\top, \ldots, {\boldsymbol{\beta}^*_{jK}}^\top)^\top,$$

$$\boldsymbol{\beta}_{jk} = (\hat{\beta}_{1jk}, \hat{\beta}_{2jk}, \hat{\beta}_{3jk})^\top, \quad \boldsymbol{\beta}^*_{jk} = (\hat{\beta}^*_{1jk}, \hat{\beta}^*_{2jk}, \hat{\beta}_{3jk})^\top,$$

$$\boldsymbol{\Omega}_j = \mathsf{COV}(\boldsymbol{\beta}_j)$$

and $\mathbf{C}$ is a contrast matrix. It is important to note that, accounting for the estimation of the equating coefficients, not only the covariance matrix of $\boldsymbol{\beta}^*_{jk}$ needs to be property calculated, but also the covariance matrices between $\boldsymbol{\beta}_{j1}$ and $\boldsymbol{\beta}^*_{jk}$ and between $\boldsymbol{\beta}^*_{jh}$ and $\boldsymbol{\beta}^*_{jk}$ are not zero. This is because the equating coefficients are estimated using the item parameter estimates obtained from group 1 and group k. For more details on the computation of the covariance matrix $\boldsymbol{\Omega}_j$ see Appendix B.

The adaptation of the test for the Rasch and the 2PL model is straightforward.

4 Simulation studies

The performance of the new proposal was assessed by means of simulation studies. Various settings were considered. The IRT models used to generate the data and estimate the item parameters are the Rasch, the 2PL and the 3PL models. The sample size for each group takes values $n = \{500, 1000, 2000, 4000, 8000\}$, while the number of items of each test is 20 and 40. Test responses of 3 groups of examinees were simulated. For each group, the θ values were generated from a normal distribution with mean $\{0, 0.5, -0.5\}$ and standard deviation $\{1, 1.2, 1.2\}$ in the 3 groups. The discrimination parameters were generated from a uniform distribution with range $[0.7, 1.3]$, the difficulty parameters were generated from a standard normal distribution, and the guessing parameters were taken equal to 0.2. The percentage of DIF items is 0%, 5% and 20%. In presence of DIF, the values added to the item parameters in the two focus groups were 0.3 and 0.5 for the discrimination parameters, and 0.4 and 0.6 for the difficulty parameters. The method used to estimate the equating coefficients is the mean-mean method (Kolen and Brennan, 2014). For each setting, 500 simulated data sets were generated. The statistical tests were applied to 2 and 3 groups (the third group was excluded when just 2 groups were considered). The traditional test was also performed for comparison. Since the proposal of this paper involves two different modifications of the traditional procedure, in order to better understand the effect of each of them, two further procedures involving only one of the two modifications were implemented. The purification procedure (Candell and Drasgow, 1988) was applied in presence of DIF items. The new and the traditional procedures were implemented in R (R Development Core Team, 2017), employing the `equateIRT` package (Battauz, 2015) for the computation of the equating coefficients. The R package `ltm` was used to fit the IRT models (Rizopoulos, 2006). In particular, the Rasch model was estimated using the function `rasch` with equal discrimination parameters, to avoid overconstraining the model. The R functions that implement the test proposed in this paper will be made publicly available in the R package [omiss].

The data sets simulated without DIF items are used to evaluate the type I error rates. The empirical type I error rates are reported in Table 1, while Figures 1, 2 and 3 give a graphical representation for the Rasch, 2PL and the 3PL models respectively. Consistently with previous studies, using the

traditional procedure, the type I error rate is lower than the significance level for the 2PL model and larger for the 3PL model.

For the Rasch model, the traditional procedure gives type I error rates lower than the significance level, while the new proposal yields type I error rates very close to the nominal level. Treating the equating coefficients as estimates (and using parameterization (1)) gives results very similar to the new proposal, while using parameterization (2) and treating the equating coefficients as known constants gives results similar to the traditional procedure.

In the case of the 2PL model (Figure 2), the type I error rates obtained with the traditional procedure are smaller than the significance level. Not reparameterizing the model improves the results obtained, but the error rates are still too low. Instead, accounting for the fact that the equating coefficients are estimates and not known constants provides error rates very close the the significance level. The new proposal, which applies both modifications, gives also very good results. For this case, the new proposal yields error rates very similar to only considering the equating coefficients as random variables. However, for small sample sizes it performs better.

In the case of the 3PL model (Figure 3), the type I error rate resulting from the traditional procedure is really huge, especially in the case of 3 groups. Considering 2 groups, avoiding reparameterization gives better results than the traditional procedure. However, increasing the sample size the type I error rate does not tend to the nominal level. In the case of 3 groups, not reparameterizing improves the performance of the test for smaller sample sizes, but gives even worse results for larger sample sizes. Considering the equating coefficients as estimates yields type I error rates closer to the significance level and the graphs show a clear trend toward the nominal level when increasing the sample size. The new proposal, in most cases, gives the closest values to the significance level. For large sample sizes and 3 groups, only considering the equating coefficients as estimates (and keeping reparameterization) provides better results, but the difference is rather small. Instead, for small sample sizes the new proposal performs definitely better than only considering the equating coefficients as estimates. Results for nominal significance levels of 0.01 and 0.10 are reported in the supplementary material, and show a very similar behavior.

[Table 1 about here.]

[Fig. 1 about here.]

[Fig. 2 about here.]

[Fig. 3 about here.]

When test responses are simulated in presence of DIF, it is also possible to evaluate the power of the test. Figures 4, 5 and 6 represent the empirical power of the tests with a percentage of 5% of DIF items. Figures 4 and 5 show that for the Rasch and the 2PL model the power of all the procedures quickly reaches 1 as the sample size increses. For the 3PL model (Figure 6), the power of the traditional procedure is higher for smaller sample sizes, and it tends to

1 as the sample size increases for all the procedures. However, it should be noted that a comparison is not appropriate since the type I error rates differ, and a greater power should be expected from a test that tends to reject the null hypothesis too often.

Results for the case of a percentage of DIF items equal to 20% are not shown because very similar to the case of a percentage of 5%.

[Fig. 4 about here.]

[Fig. 5 about here.]

[Fig. 6 about here.]

5 Discussion

In this article a new procedure to perform Wald-type tests for DIF detection is presented. The new procedure recognizes two basic aspects. One is the random nature of the estimated equating coefficients, which should be taken into account for an accurate computation of the covariance matrix. Another issue is the non-invariance to a non-linear reparametrization of the Wald test (Gregory and Veall, 1985). The simulation studies presented in this paper showed that the new proposal outperforms the traditional procedure. The results are better for the 2PL model than the 3PL model, and a sensible explanation for this difference can found in the difficulties of maximum likelihood fitting algorithms for the 3PL model (Patz and Junker, 1999). Problems with the item parameter estimation could at least partly explain the huge type I error rate, especially in small samples. On the basis of the simulation studies, the more important adjustment is given by recognizing that the equating coefficients are estimated and not known constants. For the 3PL model, even in very large sample sizes, considering the equating coefficients as constants gives very different results than considering them as estimates. This is due to the fact that in the 3PL model the correlation between the estimates of the item parameters of the same item are extremely large, and consequently the standard errors of the item parameter estimates are quite large. As a consequence, the standard errors of the equating coefficients are also large, and thus cannot be considered as constant values even in very large samples. Bayesian approaches can be used to obtain smaller standard errors (see for example Mislevy, 1986). However, further research is needed to extend the method proposed in this paper to the bayesian framework. Considering the estimates obtained from fitting the IRT model without reparameterization improves also the performance of the test for the 3PL model, especially for small sample sizes. One reason that could explain the better performance of the test with this parameterization is that is avoids the computation of the covariance matrix of the item parameters with the delta method, which introduces an approximation that is larger in small samples.

Appendix A: Equating of untransformed item parameters

Equation (12) is obtained from Equations (7) and (4) as follows:

$$\hat{\beta}^*_{2jk} = D\hat{a}^*_{jk} = \frac{D\hat{a}_{jk}}{\hat{A}_k} = \frac{\hat{\beta}_{2jk}}{\hat{A}_k}. \tag{A1}$$

Equations (7), (8) and (5) lead to Equation (13):

$$\hat{\beta}^*_{1jk} = -D\hat{a}^*_{jk}\hat{b}^*_{jk} = -D\frac{\hat{a}_{jk}}{\hat{A}_k}(\hat{A}_k\,\hat{b}_{jk}+\hat{B}_k) = -D\hat{a}_{jk}\hat{b}_{jk} - D\hat{a}_{jk}\frac{\hat{B}_k}{\hat{A}_k} = \hat{\beta}_{1jk} - \hat{\beta}_{2jk}\frac{\hat{B}_k}{\hat{A}_k}. \tag{A2}$$

Appendix B: Covariance matrix of item parameters

The covariance matrix $\boldsymbol{\Omega}_j$ entering in Equation (15) is a block matrix given by

$$\boldsymbol{\Omega}_j = \begin{pmatrix} \mathsf{COV}(\boldsymbol{\beta}_{j1}) & \mathsf{COV}(\boldsymbol{\beta}_{j1},\boldsymbol{\beta}^*_{j2}) & \mathsf{COV}(\boldsymbol{\beta}_{j1},\boldsymbol{\beta}^*_{j3}) & \dots & \mathsf{COV}(\boldsymbol{\beta}_{j1},\boldsymbol{\beta}^*_{jK}) \\ \mathsf{COV}(\boldsymbol{\beta}^*_{j2},\boldsymbol{\beta}_{j1}) & \mathsf{COV}(\boldsymbol{\beta}^*_{j2}) & \mathsf{COV}(\boldsymbol{\beta}^*_{j2},\boldsymbol{\beta}^*_{j3}) & \dots & \mathsf{COV}(\boldsymbol{\beta}^*_{j2},\boldsymbol{\beta}^*_{jK}) \\ \mathsf{COV}(\boldsymbol{\beta}^*_{j3},\boldsymbol{\beta}_{j1}) & \mathsf{COV}(\boldsymbol{\beta}^*_{j3},\boldsymbol{\beta}^*_{j2}) & \mathsf{COV}(\boldsymbol{\beta}^*_{j3}) & \dots & \mathsf{COV}(\boldsymbol{\beta}^*_{j3},\boldsymbol{\beta}^*_{jK}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathsf{COV}(\boldsymbol{\beta}^*_{jK},\boldsymbol{\beta}_{j1}) & \mathsf{COV}(\boldsymbol{\beta}^*_{jK},\boldsymbol{\beta}^*_{j2}) & \mathsf{COV}(\boldsymbol{\beta}^*_{jK},\boldsymbol{\beta}^*_{j3}) & \dots & \mathsf{COV}(\boldsymbol{\beta}^*_{jK}) \end{pmatrix}.$$

Let $\boldsymbol{\beta}_{(k)} = (\boldsymbol{\beta}_{1k}^\top,\dots,\boldsymbol{\beta}_{Jk}^\top)^\top$ denote the item parameters estimates in group k, and $\boldsymbol{\Omega}_{(k)} = \mathsf{COV}(\boldsymbol{\beta}_{(k)})$ denote the covariance matrix of the item parameter estimates in group k, which is estimated along with the estimation of the item parameters. Using the delta method, it is possible to compute the covariance matrix $\boldsymbol{\Omega} = \mathsf{COV}(\boldsymbol{\beta}_{(1)}^\top, {\boldsymbol{\beta}^*_{(2)}}^\top,\dots,{\boldsymbol{\beta}^*_{(K)}}^\top)^\top$, from which to extract $\boldsymbol{\Omega}_j$:

$$\begin{aligned}
\boldsymbol{\Omega} &= \frac{\partial(\boldsymbol{\beta}_{(1)}^\top, {\boldsymbol{\beta}^*_{(2)}}^\top,\dots,{\boldsymbol{\beta}^*_{(K)}}^\top)^\top}{\partial(\boldsymbol{\beta}_{(1)}^\top,\boldsymbol{\beta}_{(2)}^\top,\dots,\boldsymbol{\beta}_{(K)}^\top)}\mathsf{COV}((\boldsymbol{\beta}_{(1)}^\top,\boldsymbol{\beta}_{(2)}^\top,\dots,\boldsymbol{\beta}_{(K)}^\top)^\top)\frac{\partial(\boldsymbol{\beta}_{(1)}^\top, {\boldsymbol{\beta}^*_{(2)}}^\top,\dots,{\boldsymbol{\beta}^*_{(K)}}^\top)}{\partial(\boldsymbol{\beta}_{(1)}^\top,\boldsymbol{\beta}_{(2)}^\top,\dots,\boldsymbol{\beta}_{(K)}^\top)^\top} \\
&= \begin{pmatrix} \frac{\partial\boldsymbol{\beta}_{(1)}}{\partial\boldsymbol{\beta}_{(1)}^\top} & \frac{\partial\boldsymbol{\beta}_{(1)}}{\partial\boldsymbol{\beta}_{(2)}^\top} & \dots & \frac{\partial\boldsymbol{\beta}_{(1)}}{\partial\boldsymbol{\beta}_{(K)}^\top} \\ \frac{\partial\boldsymbol{\beta}^*_{(2)}}{\partial\boldsymbol{\beta}_{(1)}^\top} & \frac{\partial\boldsymbol{\beta}^*_{(2)}}{\partial\boldsymbol{\beta}_{(2)}^\top} & \dots & \frac{\partial\boldsymbol{\beta}^*_{(2)}}{\partial\boldsymbol{\beta}_{(K)}^\top} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\boldsymbol{\beta}^*_{(K)}}{\partial\boldsymbol{\beta}_{(1)}^\top} & \frac{\partial\boldsymbol{\beta}^*_{(K)}}{\partial\boldsymbol{\beta}_{(2)}^\top} & \dots & \frac{\partial\boldsymbol{\beta}^*_{(K)}}{\partial\boldsymbol{\beta}_{(K)}^\top} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Omega}_{(1)} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Omega}_{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\Omega}_{(K)} \end{pmatrix} \begin{pmatrix} \frac{\partial\boldsymbol{\beta}_{(1)}^\top}{\partial\boldsymbol{\beta}_{(1)}} & \frac{\partial\boldsymbol{\beta}^*_{(2)}{}^\top}{\partial\boldsymbol{\beta}_{(1)}} & \dots & \frac{\partial\boldsymbol{\beta}^*_{(K)}{}^\top}{\partial\boldsymbol{\beta}_{(1)}} \\ \frac{\partial\boldsymbol{\beta}_{(1)}^\top}{\partial\boldsymbol{\beta}_{(2)}} & \frac{\partial\boldsymbol{\beta}^*_{(2)}{}^\top}{\partial\boldsymbol{\beta}_{(2)}} & \dots & \frac{\partial\boldsymbol{\beta}^*_{(K)}{}^\top}{\partial\boldsymbol{\beta}_{(2)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\boldsymbol{\beta}_{(1)}^\top}{\partial\boldsymbol{\beta}_{(K)}} & \frac{\partial\boldsymbol{\beta}^*_{(2)}{}^\top}{\partial\boldsymbol{\beta}_{(K)}} & \dots & \frac{\partial\boldsymbol{\beta}^*_{(K)}{}^\top}{\partial\boldsymbol{\beta}_{(K)}} \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{\Omega}_{(1)} & \boldsymbol{\Omega}_{(1)}\frac{\partial\boldsymbol{\beta}^*_{(2)}{}^\top}{\partial\boldsymbol{\beta}_{(1)}} & \cdots & \boldsymbol{\Omega}_{(1)}\frac{\partial\boldsymbol{\beta}^*_{(K)}{}^\top}{\partial\boldsymbol{\beta}_{(1)}} \\ \frac{\partial\boldsymbol{\beta}^*_{(2)}}{\partial\boldsymbol{\beta}_{(1)}^\top}\boldsymbol{\Omega}_{(1)} & \frac{\partial\boldsymbol{\beta}^*_{(2)}}{\partial\boldsymbol{\beta}_{(1)}^\top}\boldsymbol{\Omega}_{(1)}\frac{\partial\boldsymbol{\beta}^*_{(2)}{}^\top}{\partial\boldsymbol{\beta}_{(1)}} + \frac{\partial\boldsymbol{\beta}^*_{(2)}}{\partial\boldsymbol{\beta}_{(2)}^\top}\boldsymbol{\Omega}_{(2)}\frac{\partial\boldsymbol{\beta}^*_{(2)}{}^\top}{\partial\boldsymbol{\beta}_{(2)}} & \cdots & \frac{\partial\boldsymbol{\beta}^*_{(2)}}{\partial\boldsymbol{\beta}_{(1)}^\top}\boldsymbol{\Omega}_{(1)}\frac{\partial\boldsymbol{\beta}^*_{(K)}{}^\top}{\partial\boldsymbol{\beta}_{(1)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\boldsymbol{\beta}^*_{(K)}}{\partial\boldsymbol{\beta}_{(1)}^\top}\boldsymbol{\Omega}_{(1)} & \frac{\partial\boldsymbol{\beta}^*_{(K)}}{\partial\boldsymbol{\beta}_{(1)}^\top}\boldsymbol{\Omega}_{(1)}\frac{\partial\boldsymbol{\beta}^*_{(2)}{}^\top}{\partial\boldsymbol{\beta}_{(1)}} & \cdots & \frac{\partial\boldsymbol{\beta}^*_{(K)}}{\partial\boldsymbol{\beta}_{(1)}^\top}\boldsymbol{\Omega}_{(1)}\frac{\partial\boldsymbol{\beta}^*_{(K)}{}^\top}{\partial\boldsymbol{\beta}_{(1)}} + \frac{\partial\boldsymbol{\beta}^*_{(K)}}{\partial\boldsymbol{\beta}_{(K)}^\top}\boldsymbol{\Omega}_{(K)}\frac{\partial\boldsymbol{\beta}^*_{(K)}{}^\top}{\partial\boldsymbol{\beta}_{(K)}} \end{pmatrix},
\end{aligned}$$

since $\frac{\partial\boldsymbol{\beta}_{(1)}}{\partial\boldsymbol{\beta}_{(1)}^\top}$ is the identity matrix, $\frac{\partial\boldsymbol{\beta}_{(1)}}{\partial\boldsymbol{\beta}_{(k)}^\top} = 0$ for all $k \neq 1$ and $\frac{\partial\boldsymbol{\beta}^*_{(k)}}{\partial\boldsymbol{\beta}_{(h)}^\top} = 0$ for all $h \neq k$ with $h \neq 1$. The blocks on the main diagonal of $\boldsymbol{\Omega}$ are then

$$\mathsf{COV}(\boldsymbol{\beta}^*_{(k)}) = \frac{\partial\boldsymbol{\beta}^*_{(k)}}{\partial\boldsymbol{\beta}_{(1)}^\top}\boldsymbol{\Omega}_{(1)}\frac{\partial\boldsymbol{\beta}^*_{(k)}{}^\top}{\partial\boldsymbol{\beta}_{(1)}} + \frac{\partial\boldsymbol{\beta}^*_{(k)}}{\partial\boldsymbol{\beta}_{(k)}^\top}\boldsymbol{\Omega}_{(k)}\frac{\partial\boldsymbol{\beta}^*_{(k)}{}^\top}{\partial\boldsymbol{\beta}_{(k)}},$$

while the matrices outside the main diagonal are given by

$$\mathsf{COV}(\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}^*_{(k)}) = \boldsymbol{\Omega}_{(1)} \frac{\partial \boldsymbol{\beta}^*_{(k)}}{\partial \boldsymbol{\beta}_{(1)}}^\top,$$

and

$$\mathsf{COV}(\boldsymbol{\beta}^*_{(h)}, \boldsymbol{\beta}^*_{(k)}) = \frac{\partial \boldsymbol{\beta}^*_{(h)}}{\partial \boldsymbol{\beta}^\top_{(1)}} \boldsymbol{\Omega}_{(1)} \frac{\partial \boldsymbol{\beta}^*_{(k)}}{\partial \boldsymbol{\beta}_{(1)}}^\top.$$

The chain rule can be exploited to find the derivatives

$$\frac{\partial \boldsymbol{\beta}^*_{(k)}}{\partial (\boldsymbol{\beta}^\top_{(1)}, \boldsymbol{\beta}^\top_{(k)})} = \frac{\partial \boldsymbol{\beta}^*_{(k)}}{\partial (\boldsymbol{\beta}^\top_{(k)}, \hat{A}_k, \hat{B}_k)} \frac{\partial (\boldsymbol{\beta}^\top_{(k)}, \hat{A}_k, \hat{B}_k)^\top}{\partial (\boldsymbol{\beta}^\top_{(1)}, \boldsymbol{\beta}^\top_{(k)})}, \tag{B1}$$

where

$$\frac{\partial (\hat{A}_k, \hat{B}_k)^\top}{\partial (\boldsymbol{\beta}^\top_{(1)}, \boldsymbol{\beta}^\top_{(k)})} = \frac{\partial (\hat{A}_k, \hat{B}_k)^\top}{\partial (\mathbf{v}^\top_{(1)}, \mathbf{v}^\top_{(k)})} \frac{\partial (\mathbf{v}^\top_{(1)}, \mathbf{v}^\top_{(k)})^\top}{\partial (\boldsymbol{\beta}^\top_{(1)}, \boldsymbol{\beta}^\top_{(k)})}, \tag{B2}$$

where $\mathbf{v}_{(k)} = (\mathbf{v}^\top_{1k}, \ldots, \mathbf{v}^\top_{Jk})^\top$. The non-zero derivatives entering in (B1) and (B2) are given in the following (derivatives of a variable with respect to itself are not shown):

$$\frac{\partial \hat{\beta}^*_{1jk}}{\partial \hat{\beta}_{1jk}} = 1, \quad \frac{\partial \hat{\beta}^*_{1jk}}{\partial \hat{\beta}_{2jk}} = -\frac{\hat{B}_k}{\hat{A}_k}, \quad \frac{\partial \hat{\beta}^*_{1jk}}{\partial \hat{A}_k} = \hat{\beta}_{2jk} \frac{\hat{B}_k}{\hat{A}^2_k},$$

$$\frac{\partial \hat{\beta}^*_{1jk}}{\partial \hat{B}_k} = -\frac{\hat{\beta}_{2jk}}{\hat{A}_k}, \quad \frac{\partial \hat{\beta}^*_{2jk}}{\partial \hat{\beta}_{2jk}} = \frac{1}{\hat{A}_k}, \quad \frac{\partial \hat{\beta}^*_{2jk}}{\partial \hat{A}_k} = -\frac{\hat{\beta}_{2jk}}{\hat{A}^2_k}$$

$$\frac{\partial \hat{a}_{jk}}{\partial \hat{\beta}_{2jk}} = \frac{1}{D}, \quad \frac{\partial \hat{b}_{jk}}{\partial \hat{\beta}_{1jk}} = -\frac{1}{\hat{\beta}_{2j1}}, \quad \frac{\partial \hat{b}_{jk}}{\partial \hat{\beta}_{2jk}} = \frac{\hat{\beta}_{1jk}}{\hat{\beta}^2_{2jk}}.$$

The derivatives $\frac{\partial (\hat{A}_k, \hat{B}_k)^\top}{\partial (\mathbf{v}^\top_{(1)}, \mathbf{v}^\top_{(k)})}$ are given in Ogasawara (2000) and Ogasawara (2001).

References

Bartholomew DJ, Knott M, Moustaki I (2011) Latent variable models and factor analysis: A unified approach. West Sussex: John Wiley & Sons

Battauz M (2015) equateIRT: An R package for IRT test equating. Journal of Statistical Software 68(7):1–22, doi=10.18637/jss.v068.i07

Bock RD, Aitkin M (1981) Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika 46(4):443–459, doi: 10.1007/BF02293801

Candell GL, Drasgow F (1988) An iterative procedure for linking metrics and assessing item bias in item response theory. Applied psychological measurement 12(3):253–260, doi: 10.1177/014662168801200304

Casella G, Berger RL (2002) Statistical inference. Pacific Grove: Duxbury

Gregory AW, Veall MR (1985) Formulating Wald tests of nonlinear restrictions. Econometrica 53(6):1465–1468, doi: 10.2307/1913221

Kim SH, Cohen AS, Kim HO (1994) An investigation of Lord's procedure for the detection of differential item functioning. Applied Psychological Measurement 18(3):217–228, doi: 10.1177/014662169401800303

Kim SH, Cohen AS, Park TH (1995) Detection of differential item functioning in multiple groups. Journal of Educational Measurement 32(3):261–276, doi: 10.1111/j.1745-3984.1995.tb00466.x

Kolen M, Brennan R (2014) Test Equating, Scaling, and Linking: Methods and Practices (3rd ed.). New York: Springer

van der Linden W (2016) Handbook of Item Response Theory, Volume One: Models. Boca Raton: Chapman & Hall/CRC

Lord FM (1980) Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum

Magis D, Béland S, Tuerlinckx F, De Boeck P (2010) A general framework and an R package for the detection of dichotomous differential item functioning. Behavior Research Methods 42(3):847–862, doi:10.3758/BRM.42.3.847

Mislevy RJ (1986) Bayes modal estimation in item response models. Psychometrika 51(2):177–195, DOI 10.1007/BF02293979

Ogasawara H (2000) Asymptotic standard errors of IRT equating coefficients using moments. Economic Review (Otaru University of Commerce) 51(1):1–23

Ogasawara H (2001) Standard errors of item response theory equating/linking by response function methods. Applied Psychological Measurement 25(1):53–67, doi: 10.1177/01466216010251004

Patz RJ, Junker BW (1999) Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. Journal of educational and behavioral statistics 24(4):342–366, doi: 10.3102/10769986024004342

R Development Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/, ISBN 3-900051-07-0

Reise SP, Revicki DA (2014) Handbook of item response theory modeling: Applications to typical performance assessment. New York: Routledge

Rizopoulos D (2006) ltm: An R package for latent variable modeling and item response theory analyses. Journal of statistical software 17(5):1–25, doi:10.18637/jss.v017.i05

List of Figures

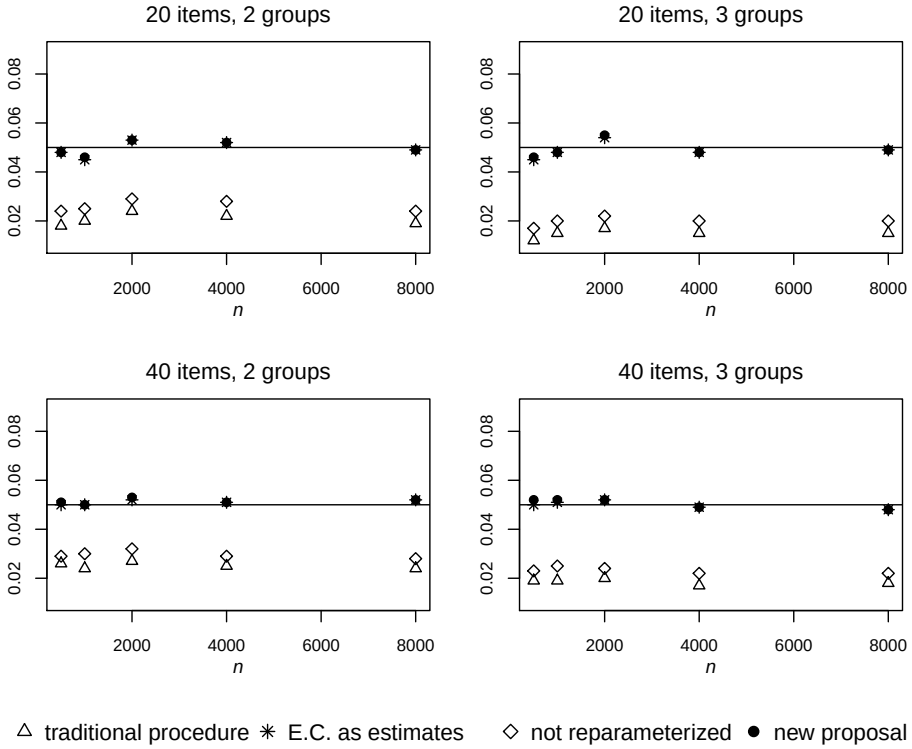


Fig. 1 Type I error rates (false positive rate) for the 1PL model.

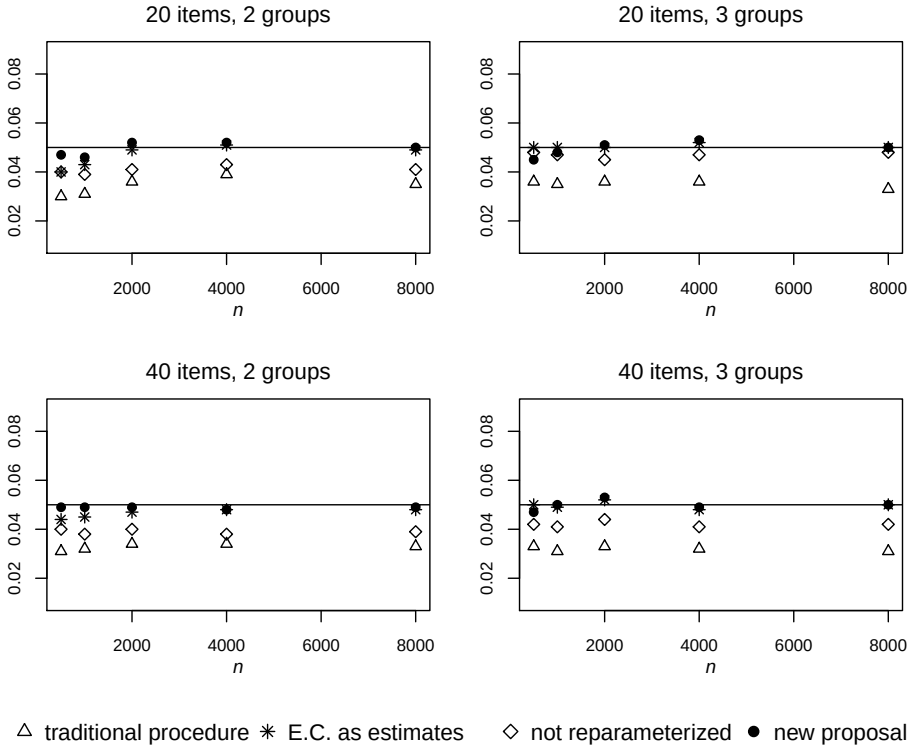


Fig. 2 Type I error rates (false positive rate) for the 2PL model.

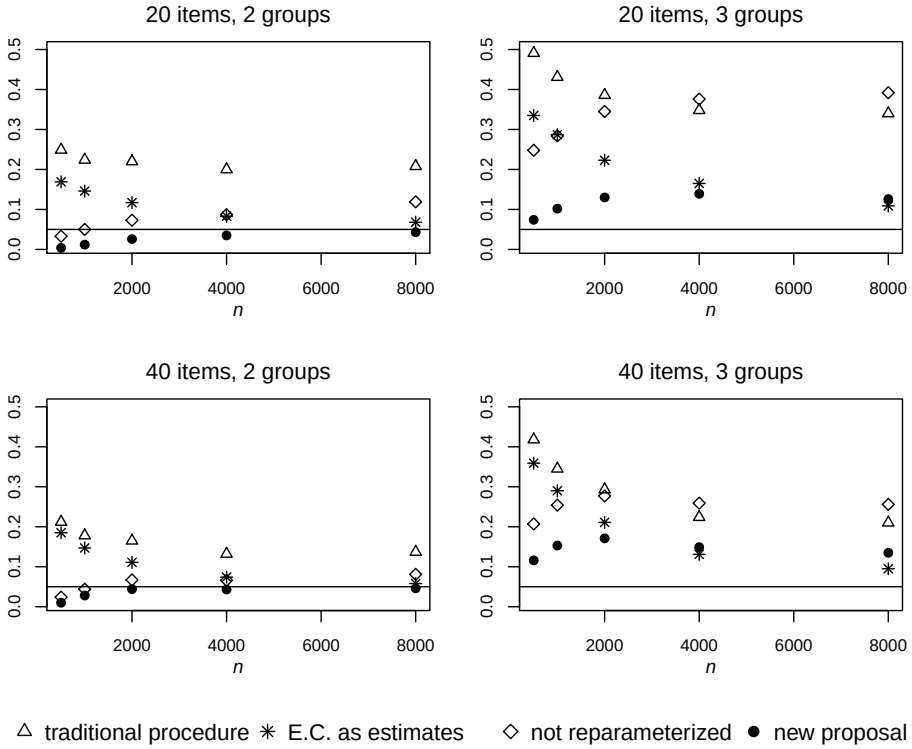
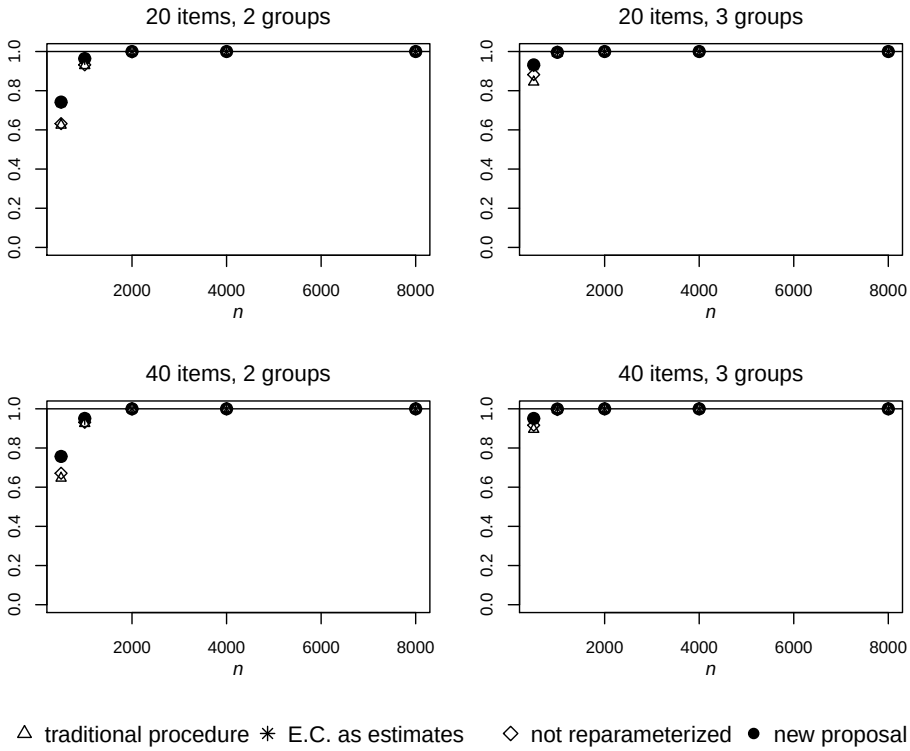


Fig. 3 Type I error rates (false positive rate) for the 3PL model.



△ traditional procedure ✳ E.C. as estimates ◇ not reparameterized ● new proposal

Fig. 4 Power (true positive rate) for the Rasch model (percentage of DIF items: 5%.).

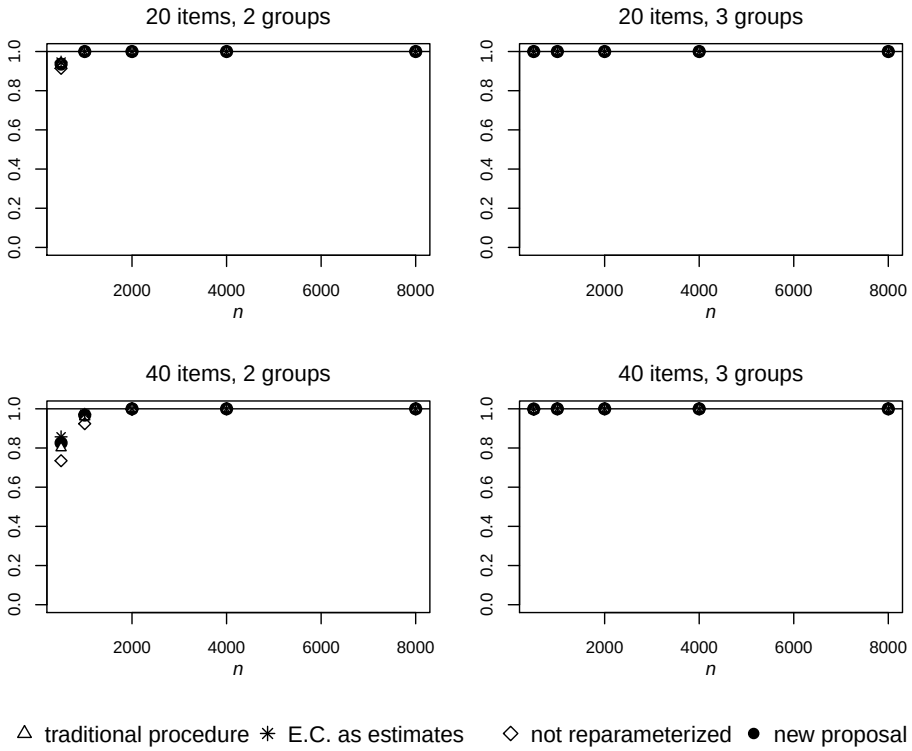


Fig. 5 Power (true positive rate) for the 2PL model (percentage of DIF items: 5%.).

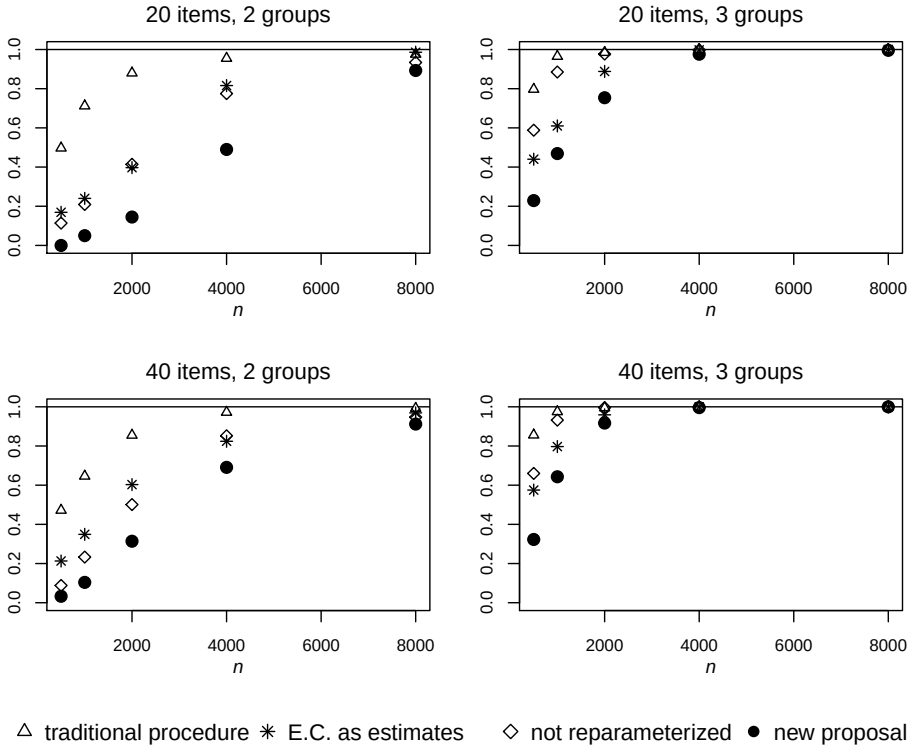


Fig. 6 Power (true positive rate) for the 3PL model (percentage of DIF items: 5%.).

List of Tables

Table 1 Type I error rate (false positive rate).

			2 groups				3 groups			
model	items	n	traditional	E.C. as est.	not repar.	new	traditional	E.C. as est.	not repar.	new
1PL	20	0500	0.018	0.048	0.024	0.048	0.012	0.045	0.017	0.046
1PL	20	1000	0.020	0.045	0.025	0.046	0.015	0.048	0.020	0.048
1PL	20	2000	0.024	0.053	0.029	0.053	0.017	0.054	0.022	0.055
1PL	20	4000	0.022	0.052	0.028	0.052	0.015	0.048	0.020	0.048
1PL	20	8000	0.019	0.049	0.024	0.049	0.015	0.049	0.020	0.049
1PL	40	0500	0.026	0.050	0.029	0.051	0.019	0.050	0.023	0.052
1PL	40	1000	0.024	0.050	0.030	0.050	0.019	0.051	0.025	0.052
1PL	40	2000	0.027	0.052	0.032	0.053	0.020	0.052	0.024	0.052
1PL	40	4000	0.025	0.051	0.029	0.051	0.017	0.049	0.022	0.049
1PL	40	8000	0.024	0.052	0.028	0.052	0.018	0.048	0.022	0.048
2PL	20	500	0.030	0.040	0.040	0.047	0.036	0.050	0.048	0.045
2PL	20	1000	0.031	0.043	0.039	0.046	0.035	0.050	0.047	0.048
2PL	20	2000	0.036	0.049	0.041	0.052	0.036	0.050	0.045	0.051
2PL	20	4000	0.039	0.051	0.043	0.052	0.036	0.052	0.047	0.053
2PL	20	8000	0.035	0.049	0.041	0.050	0.033	0.050	0.048	0.050
2PL	40	500	0.031	0.044	0.040	0.049	0.033	0.050	0.042	0.047
2PL	40	1000	0.032	0.045	0.038	0.049	0.031	0.049	0.041	0.050
2PL	40	2000	0.034	0.047	0.040	0.049	0.033	0.052	0.044	0.053
2PL	40	4000	0.034	0.048	0.038	0.048	0.032	0.048	0.041	0.049
2PL	40	8000	0.033	0.048	0.039	0.049	0.031	0.050	0.042	0.050
3PL	20	500	0.249	0.169	0.033	0.004	0.491	0.335	0.248	0.074
3PL	20	1000	0.224	0.146	0.050	0.012	0.431	0.287	0.284	0.102
3PL	20	2000	0.220	0.117	0.073	0.026	0.386	0.223	0.345	0.130
3PL	20	4000	0.200	0.083	0.087	0.035	0.348	0.165	0.376	0.139
3PL	20	8000	0.208	0.068	0.119	0.043	0.340	0.109	0.392	0.126
3PL	40	500	0.212	0.185	0.024	0.010	0.418	0.359	0.207	0.116
3PL	40	1000	0.178	0.147	0.044	0.028	0.345	0.290	0.254	0.153
3PL	40	2000	0.165	0.111	0.067	0.044	0.293	0.211	0.277	0.171
3PL	40	4000	0.132	0.074	0.066	0.043	0.224	0.131	0.259	0.149
3PL	40	8000	0.137	0.058	0.081	0.046	0.210	0.095	0.256	0.135